

This application is submitted in the name of the following inventor(s):

<i>Inventor</i>	<i>Citizenship</i>	<i>Residence City and State</i>
Srinivasan VISWANATHAN	India	Fremont, California
Steven R. KLEIMAN	United States	Los Altos, California

The assignee is *Network Appliance, Inc.*, a corporation having an office at 495 East Java Drive, Sunnyvale California 94089.

## TITLE OF THE INVENTION

Recovery of File System Data in File Servers Mirrored File System Volumes

## BACKGROUND OF THE INVENTION

### *1. Field of the Invention*

The invention relates to recovery of file system data in file servers having mirrored file system volumes.

## 2. *Related Art*

Network file servers and other file systems are subject to errors and other failures, including those arising from hardware failure, software error, or erroneous configuration. Because of the possibility of error, many file systems provide additional copies of data in the file system, such as by providing a mirrored file system volume. In a mirrored file system, a first volume provides a first copy of the file system, while a second volume provides a synchronous, second copy of the file system. Thus, if data on the first volume is corrupted or otherwise lost, data from the second volume can be used in its place transparently.

One problem in the known art is that the first volume and second volume of the file system can fail to remain in synchronization. Thus, each volume of the mirrored file system would include a set of files or other objects from a different timestamp (or checkpoint) in the file system history. As a result, the first volume and second volume will no longer serve as accurate mirrors for each other because one is out-of-date. An aspect of this problem is that, after system crashes, it is unknown which of the first volume and second volume is the most recent. Accordingly, it would be desirable to assure that the first volume and second volume of the file system remain synchronized after system crashes. If it is not possible for the first volume and second volume to remain synchronized, it is desirable to rapidly determine which is the most recent version and use efficiently, so as to cause resynchronization.

1           A first known method is to resynchronize the two mirror copies after system  
2 crashes by copying every block. While this method can generally achieve the result of as-  
3 suring that the first copy and second copy of the file system are synchronized after system  
4 crashes, it has the severe drawback that it is very inefficient, as each file block of at least  
5 one of the mirror file systems must be copied to the other one of the mirror file systems.  
6 When the file system is particularly large, such as one that approaches or exceeds a tera-  
7 byte in size, this drawback makes this known method untenable due to its incredible la-  
8 tency (and liability to other failures).

9  
10           A second known method is to maintain a log of regions or file blocks in  
11 each mirrored volume that have been changed (sometimes known as “dirty” file blocks).  
12 When such a log is maintained, it is only necessary to copy those file blocks that are dirty,  
13 rather than an entire mirrored volume. While this method can generally achieve the result  
14 otherwise achieved by the first known method, is still subject to at least two drawbacks.  
15 First, this method is more complex, in that it requires careful maintenance so as to ensure  
16 that the log remains synchronous. Second, the log itself must generally be mirrored for  
17 reliability, which of course re introduces the entire problem of recovery of mirrored files  
18 after system crashes. Third, maintaining this additional log increases the latency of every  
19 operation. Moreover, such a technique can introduce additional errors in the event that  
20 the log is unreliable.

1           Accordingly, it would be desirable to provide a technique for recovery of  
2 file system data in file servers having mirrored file system volumes that is not subject to  
3 drawbacks of the known art.

## 4 5 SUMMARY OF THE INVENTION

6  
7           The invention provides a method and system for recovery of file system  
8 data in file servers having mirrored file system volumes. In a preferred embodiment, the  
9 invention makes use of a consistency point model including a snapshot feature of a robust  
10 file system (the "WAFL File System"), such as disclosed in the Incorporated Disclosures,  
11 to rapidly determine which of two or more mirrored volumes is most up-to-date, and  
12 which blocks of the most recent mirrored volume have been changed from each one of  
13 the mirrored file systems. Among a plurality of two or more mirrored volumes, the in-  
14 vention rapidly determines which is the most up-to-date by examining a most recent con-  
15 sistency point number maintained by the WAFL File System at each mirrored volume.  
16 The invention rapidly and reliably determines what blocks are shared between that most  
17 up-to-date mirrored volume and each other mirrored volume, in response to a snapshot of  
18 the file system maintained at each mirrored volume and are stored in common pairwise  
19 between each mirrored volume and the most up-to-date mirrored volume. The invention  
20 copies only those blocks that have been changed between the common snapshot and the  
21 most up-to-date snapshot. This rapid and reliable comparison of blocks, followed by the

1 efficient transfer of those blocks that have been changed, does not present drawbacks of  
2 the known art.

3  
4 The invention provides an enabling technology for a wide variety of appli-  
5 cations for file system recovery using redundant file systems, so as to obtain substantial  
6 advantages and capabilities that are novel and non-obvious in view of the known art. Ex-  
7 amples described below primarily relate to mirrored file system volumes in a network file  
8 server, but the invention is broadly applicable to many different types of redundant file  
9 systems, such as those used in RAID subsystems and parallel storage systems.

#### 11 BRIEF DESCRIPTION OF THE DRAWINGS

12  
13 Figure 1 shows a block diagram of a system for recovery of file system data  
14 in file servers having mirrored file system volumes.

15  
16 Figure 2 shows a process flow diagram of a method for operating a system  
17 as in figure 1.

#### 19 DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

20  
21 In the following description, a preferred embodiment of the invention is de-  
22 scribed with regard to preferred process steps and data structures. Embodiments of the

1 invention can be implemented using general-purpose processors or special purpose proc-  
2 essors operating under program control, or other circuits, adapted to particular process  
3 steps and data structures described herein. Implementation of the process steps and data  
4 structures described herein would not require undue experimentation or further invention.

5  
6 *Lexicography*

7  
8 The following terms refer or relate to aspects of the invention as described  
9 below. The descriptions of general meanings of these terms are not intended to be limit-  
10 ing, only illustrative.

- 11
- 12 • *block* — in general, any collection of data for data objects in a file system.
  - 13
  - 14 • *consistency point* — in general, any point at which the consistency of a file system  
15 is assured or recorded.
  - 16
  - 17 *file server* — in general, any device which responds to messages requesting file  
18 system operations.
  - 19
  - 20 • *file system* — in general, any organization or structure of information for storage  
21 or retrieval.
  - 22

- 1 • *file system data* — in general, any information recorded in a file system or an ob-  
2 ject in a file system.  
3
- 4 • *file system volume* — in general, any mass storage device, or collection thereof, for  
5 storage or retrieval of file system objects.  
6
- 7 *mirrored volume* — in general, any file system volume having a copy of at least a  
8 portion of another file system volume.  
9
- 10 *parallel storage system* — in general, any file system in which data is recorded, in  
11 whole or in part, in multiple locations or multiple ways.  
12
- 13 *RAID subsystem* — in general, any system including a redundant array of mass  
14 storage drives.  
15
- 16 *recovery of file system data* — in general, any recopying or regeneration of infor-  
17 mation from one memory or storage medium to another.  
18
- 19 *redundant file system* — in general, any file system in which data is recorded, in  
20 whole or in part, with additional information allowing the recovery of at least a  
21 portion of that data.  
22

1        *re-synchronize* — in general, any operation in which objects in a file system are  
2        reorganized or rewritten to assure that file system objects maintain or restore syn-  
3        chronization.

4  
5        *shared file block* — in general, any file block whose data contents are located on  
6        more than one file system volume.

7  
8        *snapshot* — in general, any consistent file system available, in whole or in part, for  
9        later retrieval even if the snapshot is not a current consistent file system.

10  
11       *up-to-date* — in general, a measure of recentness of a file system, file system ob-  
12       ject, or snapshot.

13  
14       *WAFL File System* — in general, a file system as described in the Incorporated  
15       Disclosures, or any file system in which at least one snapshot is maintained in ad-  
16       dition to a current consistent file system.

17  
18                As noted above, these descriptions of general meanings of these terms are  
19       not intended to be limiting, only illustrative. Other and further applications of the inven-  
20       tion, including extensions of these terms and concepts, would be clear to those of ordinary  
21       skill in the art after perusing this application. These other and further applications are part



of the scope and spirit of the invention, and would be clear to those of ordinary skill in the art, without further invention or undue experimentation.

#### *System Elements*

Figure 1 shows a block diagram of a system for recovery of file system data in file servers having mirrored file system volumes.

A system 100 includes a file server (or other device) 110, a communication network 120, and a network interface 130. The file server 110 includes a plurality of mirrored file system volumes 111, each of which includes mass storage for recording and retrieving data. Each file system volume 111 includes at least one snapshot 112 according to the WAFL File System, as described in the Incorporated Disclosures. Each snapshot 112 includes a file system information block 113, including a pointer to an entire consistent file system and a consistency point value 114 indicating a sequence in which that snapshot 112 was generated.

Each file system volume 111 also includes an active file system 115, itself associated with a consistent point value 114. In a preferred embodiment, snapshots 112 are made periodically in response to (and as copies of) an active file system 115. Thus, while every snapshot 112 includes a consistent point value 114 from its associated active

1 file system 115, not every active file system 115 is made into a snapshot, and thus not  
2 every consistency point value 114 is associated with a snapshot 112.

3  
4 The file server 110 receives messages 116 requesting to write data or otherwise alter data from the communication network 120 using the network interface 130.  
5  
6 In normal operation, the file server 110 parses those messages 116 and writes the same  
7 data to both of the active file systems 115 of the mirrored file system volumes 111, so that  
8 each of the mirrored file system volumes 111 includes the same active file systems 115,  
9 the same snapshots 112, therefore the same data. However, in the event of a system crash  
10 or other error, it might occur that one or more of the mirrored file system volumes 111  
11 fails to remain in synchronization with the others, either because its active file system 115  
12 is not up-to-date or its snapshots 112 are not up-to-date.

13  
14 If one or more of the mirrored file system volumes 111 is not in synchroni-  
15 zation with the others, there will be at least one mirrored file system volume 111 having  
16 an active file system 115 with a consistency point value 114 larger than all others. This  
17 indicates that the associated an active file system 115 and the associated file system vol-  
18 ume 111 (with the highest consistency point value 114) is the most up-to-date file system  
19 volume 111 of all of the mirrored file system volumes 111.

20  
21 Similarly, for any pair of mirrored file system volumes 111, there will be at  
22 least one common snapshot 112 present for them both, thus having the same consistency

1 point value 114 for the common snapshot 112 at each of the two mirrored file system vol-  
2 umes 111. For any pair of mirrored file system volumes 111 A and B, the difference be-  
3 tween the common snapshot 112 and the most up-to-date active file system 115 (say, at  
4 mirrored file system volume 111 A) can be easily and rapidly determined using the  
5 WAFL File System. The file blocks indicated by that difference are the only file blocks  
6 necessary for re-synchronization between the pair of mirrored file system volumes 111 A  
7 and B.

8  
9 While each pair (A and B) of mirrored file system volumes 111 will have at  
10 least one common snapshot 112, of which one can be compared with the most up-to-date  
11 active file system 115, there is no particular requirement that each other pair (A and C, or  
12 A and D) of mirrored file system volumes 111 will have the same common snapshot 112  
13 as the first such pair (A and B). However, for each such other pair (A and C, or A and D)  
14 of mirrored file system volumes 111, the difference between the common snapshot 112  
15 and the most up-to-date active file system 115 can still be easily and rapidly determined  
16 using the WAFL File System; the file blocks indicated by that difference are the only file  
17 blocks necessary for re-synchronization between the other pair (A and C, or A and D) of  
18 mirrored file system volumes 111.

## 1 *Method of Operation*

2  
3 Figure 2 shows a process flow diagram of a method for operating a system  
4 as in figure 1.

5  
6 A method 200 includes a set of flow points and a set of steps. The system  
7 100 performs the method 200. Although the method 200 is described serially, the steps of  
8 the method 200 can be performed by separate elements in conjunction or in parallel,  
9 whether asynchronously, in a pipelined manner, or otherwise. There is no particular re-  
10 quirement that the method 200 be performed in the same order in which this description  
11 lists the steps, except where so indicated.

12  
13 At a flow point 210, the file server 110 is ready to re-synchronize a plurality  
14 of mirrored file system volumes 111.

15  
16 At a step 211, the file server 110 examines the file system information  
17 block 113 for each one of the plurality of mirrored file system volumes 111, to determine  
18 a single consistency point value 114 which is the maximum for all active file systems 115  
19 at such mirrored file system volumes 111. While it is possible that there will be more  
20 than one such mirrored file system volume 111 having an active file system 115 with that  
21 maximum consistency point value 114, there is no particular requirement to select one of

1 such mirrored file system volumes 111 in preference to others, as all active file systems  
2 115 with that identical consistency point value 114 will be identical.

3  
4 At a step 212, the mirrored file system volumes 111 with the maximum  
5 consistency point value 114 for an active file system 115 generates a new snapshot 112  
6 for that active file system 115 and having that maximum consistency point value 114.  
7 This new snapshot 112 is thus the most up-to-date snapshot 112 and has the maximum  
8 consistency point value 114.

9  
10 At a step 213, for each one of the plurality of mirrored file system volumes  
11 111 (other than the file system volumes 111 with the most up-to-date active file system  
12 115) the file server 110 examines the file system information block 113, to determine a  
13 snapshot 112 at that one mirrored file system volume 111 that is common with the mir-  
14 rored file system volume 111 having the most up-to-date snapshot 112. Thus, the file  
15 server 110 determines a closest degree of synchronization between each mirrored file  
16 system volume 111 (in turn) and the mirrored file system volume 111 having the most up-  
17 to-date snapshot 112.

18  
19 At a step 214, for each such closest degree of synchronization, the file  
20 server 110 determines a difference between the common snapshot 112 and the most up-  
21 to-date snapshot 112, thus generating a set of file blocks that have been changed between  
22 the common snapshot 112 and the most up-to-date snapshot 112. These changed file

1 blocks are the only file blocks required to be re-synchronized between the common snap-  
2 shot 112 and the most up-to-date active file system 115.

3  
4 At a step 215, for each such set of changed file blocks, the file server 110  
5 re-synchronizes each mirrored file system volume 111 with the most up-to-date snapshot  
6 112 by copying only the changed file blocks over, thus generating a copy of the most up-  
7 to-date snapshot 112 at each mirrored file system volume 111.

8  
9 In a preferred embodiment, there are only two such mirrored file system  
10 volumes 111. The file server 110 needs to make only one comparison to determine the  
11 maximum consistency point value 114 for a most up-to-date active file system 115. The  
12 file server 110 needs to examine only one pair of mirrored file system volumes 111 for a  
13 common snapshot 112. The file server 110 needs to determine only one set of changed  
14 blocks between the common snapshot 112 and the most up-to-date snapshot 112. The file  
15 server 110 needs to copy only one set of changed blocks from one mirrored file system  
16 volume 111 to the other.

17  
18 However, in alternative embodiments, there may be more than two mirrored  
19 file system volumes 111. Those skilled in the art will see, after perusal of this applica-  
20 tion, that the invention is easily and readily generalized to additional mirrored file system  
21 volumes 111, without undue experimentation or further invention.

1           In a preferred embodiment, the mirrored file system volumes 111 can each  
2 be updated to create new active file systems 115 in response to messages 116 requesting  
3 file system operations, even while the snapshot 112 at each mirrored file system volumes  
4 111 is being synchronized with the most up-to-date snapshot 112. Thus, the mirrored file  
5 system volumes 111 can each perform the full functions of a file server 110 mirrored file  
6 system volume 111 even while the re-synchronization is taking place.

7  
8           After this step, the method 200 has re-synchronized all of the mirrored file  
9 system volumes 111 to the most up-to-date active file system 115.

10  
11           In a preferred embodiment, the method 200 is performed each time the sys-  
12 tem 100 recovers from a system crash, as part of the crash recovery process. In alterna-  
13 tive embodiments, the method 200 may be performed in response to other events, such as  
14 in response to a timer, in response to detection of lack of synchronization between the  
15 mirrored volumes, or in response to operator command.

### 16 17 *Generality of the Invention*

18  
19           The invention has general applicability to various fields of use, not neces-  
20 sarily related to the services described above. For example, these fields of use can in-  
21 clude one or more of, or some combination of, the following:

- 1 • file system recovery using redundant file systems other than mirrored file system
- 2 volumes
- 3
- 4 • RAID subsystems
- 5
- 6 parallel storage systems
- 7

8 Other and further applications of the invention in its most general form, will  
9 be clear to those skilled in the art after perusal of this application, and are within the  
10 scope and spirit of the invention. Although preferred embodiments are disclosed herein,  
11 many variations are possible which remain within the concept, scope, and spirit of the in-  
12 vention, and these variations would become clear to those skilled in the art after perusal  
13 of this application.